# EDRN Informatics Center Report
# 2005-2011

Principal Investigator: Daniel J. Crichton
NASA Jet Propulsion Laboratory

August 2011

# EXECUTIVE SUMMARY

The Early Detection Research Network (EDRN) of the U.S. National Cancer Institute (NCI) consisting of a research network of collaborating scientists from over 40 institutions is focused on identifying and validating cancer biomarkers at their earliest stages. “The work of the EDRN is concentrated on … the discovery of markers, the validation of markers in distinguishing the presence or absence of cancer, and testing markers for the ability to detect preclinical and early-stage disease” [1]. The EDRN was founded in 2000 and is currently led by the Cancer Biomarkers Research Group, the Division of Cancer Prevention, National Cancer Institute. It is comprised of Biomarker Discovery Laboratories, Biomarker Reference Laboratories, Clinical and Epidemiological Centers, and a Data Management and Coordinating Center. The Jet Propulsion Laboratory serves as the informatics center. For details, see www.cancer.gov/edrn.

EDRN’s core mission is the discovery and validation of biomarkers. Critical to that mission is having an informatics infrastructure that is “biomarker-centric”. This involves having a core ontology model that describes the elements of biomarker research that span the entire spectrum of activities of the EDRN investigators i.e. from tissue banking, to managing information about proposed biomarkers, to validation studies. Capturing the information into a “virtual data grid” allows EDRN to construct the EDRN Knowledge System which serves as an online, distributed resource of data and information described in EDRN’s core ontology. The EDRN Knowledge System promises to dramatically improve the capability for scientific research by enabling real-time access to a variety of information that crosses institutional boundaries. There are clear scenarios for how such a system can improve the discovery process, flexibility and agility are at the core given the dynamic nature of cancer biomarker research.

The EDRN Knowledge System is built on the notion of a “data grid”. The data grid allows for linking loosely related items across a highly heterogeneous, distributed environment. The infrastructure of the data grid is a set of information system services that allow distributed databases to be integrated and accessed real-time [2,3]. EDRN has been a pioneer at developing the “data grid” concept for bioinformatics having partnered with NASA’s Jet Propulsion Laboratory to deploy JPL’s data grid software tools for virtualized access to biospecimen data repositories distributed at several EDRN sites in 2002.

Recognizing the need to build an effective knowledge system where biospecimens, scientific data, study specific data, and biomarker data can be captured, accessed, and shared at a national level via a transparent, grid-type architecture, the EDRN has focused on addressing five informatics goals:

1. defining an information model (in the form of an ontology) for describing the EDRN problem space;
2. enabling all components of the knowledge system to be distributed;
3. providing software interfaces for capture, discovery, and access of data resources across the knowledge system;
4. providing a secure transfer and distribution infrastructure to meet United States Federal regulations for data sharing of health information; and
5. providing an integrated portal environment for access to the distributed EDRN data.

The EDRN Knowledge System architecture takes a very pragmatic approach by deconstructing the process of biomarker research in early detection into a set of functions and providing a layered system with applications constructed on top of the infrastructure to enable the EDRN, as a collection of research institutions, to be integrated. The applications represent the critical functions that are performed by the research community in biomarker research. Furthermore, by integrating these applications into an enterprise system, the EDRN is able to provide the capability for managing the biomarker information assets at a national level. The architecture is therefore decomposed into a set of projects that make up the informatics portfolio. These projects are implemented across the EDRN by a diverse informatics team located at multiple research institutions.

These projects include:

1. development of Common Data Elements (CDEs) to explicitly capture and manage data attributes in a consistent manner;
2. development of a core Biomarker Ontology, organizing the CDEs into a set of objects and relationships that represents the information space of biomarker research;
3. development of an integrated system for accessing and sharing biospecimen information including specimen, epidemiological, and study information from biomarker research. EDRN titled this project the "EDRN Resource Network Exchange" or ERNE;
4. development of a Biomarker Database for annotating information about biomarkers and their relationship to studies and publications;
5. development of an information system for Study Management. This includes support for EDRN validation studies and provides an infrastructure for capturing and managing the information about EDRN studies;
6. development of an infrastructure for capturing and warehousing results. These results include the raw and processed data captured from EDRN validation studies. EDRN provides a common software component called the EDRN Catalog and Archive System (eCAS) which can be configured to capture information across very different validation studies. This information is then integrated into the EDRN Knowledge System;
7. development of a Biomarker Atlas which allows for researchers to search the biomarker database based on anatomical maps (lung, colon, breast, etc). The Biomarker Atlas allows for virtualized access to the information captured above in ERNE, eCAS, the Biomarker Database and the EDRN Study Management System (VSIMS); and
8. development of a Public Portal for sharing EDRN data results with the research community. The EDRN Public Portal allows for access to all of the EDRNs published information. It provides a "Google-like" search capability for searching the EDRN Knowledge System allowing users to browse and access the data assets. The EDRN Public Portal provides multi-level security architecture to allow role-based access to information. This allows EDRN to target releases of information for certain communities without compromising access to sensitive data.

JPL has formed and led the Informatics Center at the Jet Propulsion Laboratory helping EDRN become a pioneer in cancer biomarker research and informatics. As JPL has shown in other science domains from exploring the universe to monitoring the earth through observation and science research, JPL has helped to leverage informatics technology to support scientific research enabling new approaches to scientific discovery. Advances in information technology are paving the way for scientific research networks to be constructed so that integrated and collaborative studies can be performed. More and more, modeling and distributed system technologies are allowing for enterprise systems which provides access to distributed data and computational resources as an integrated "grid" of information and services. "Clearly, virtualized grids are in their infancy, but the needs of programs like the EDRN are demonstrating the benefits and the criticality for bringing scientific research endeavors together into a secure, integrated enterprise to support collaboration and discovery" [2] of biomarkers. The EDRN is positioned to provide such a capability at a national level for biomarker research, discovery and validation.

## I. INTRODUCTION

The EDRN program is established to: (a) promote translational research to identify biomarkers for cancer risk, early detection, and molecular diagnosis and prognosis of early cancer; and (b) coordinate biomarker research and therapeutic strategies in order to reduce cancer morbidity and mortality. These goals are to be achieved through strategic and systematic, evidence-based discovery, development, and validation of biomarkers (for details, see the EDRN Objectives in the Guidelines document on the EDRN web site at http://edrn.nci.nih.gov/FOA-guidelines).

The EDRN is comprised of several components including Biomarker Development Laboratories, Biomarker Reference Laboratories, Clinical Epidemiology and Validation Centers, the Data Management and Coordinating Center and the Informatics Center. The Biomarker Development Laboratories (BDL) are responsible for the development and characterization of new biomarkers or the refinement of existing biomarkers. The Biomarker Reference Laboratories (BRL) serve as a Network resource for the clinical and laboratory validation of biomarkers, which includes technological development, quality control, refinement, and high throughput. The Clinical Validation Centers (CVC) conduct clinical and epidemiological research regarding the clinical application of biomarkers. The Data Management and Coordination Center (DMCC) coordinates the EDRN research activities, providing logistic support, and conducting statistical and computational research for data analysis, analyzing data for validation. The Informatics Center supports EDRN's efforts through the development of a Network-wide informatics infrastructure to support data access, sharing and discovery.

In the area of informatics, the EDRN has been a pioneer in providing a strong informatics component to support biomarker research using novel approaches. As a distributed, collaborative network, EDRN relies on a strong infrastructure to support research across cancer research institutions and individual laboratories. Since its inception, EDRN has had a grand vision for a knowledge system that enables researchers the ability to capture, access and share study and results from biomarker research using a modern informatics infrastructure.

Since 2001, JPL has been working with the EDRN through an interagency agreement that allowed for the transfer and infusion of data and informatics system technology from space research. This has quickly allowed the EDRN to construct virtual systems that allow for sharing of data at a national level. In 2005, JPL became the "Informatics Center" for EDRN and has worked to ensure that the informatics systems within EDRN are tightly integrated to support biomarker research using innovative approaches for constructing such a virtual system.

The virtual system is unique across the National Cancer Institute by bridging diverse systems in order to support scientific collaboration and increase the ability to perform analysis of data across multiple computing environments. The EDRN Knowledge Environment, as it is known, is the flagship activity of the EDRN informatics team. It is a national data grid that allows diverse systems to be integrated and to share data for the EDRN. In traditional information systems, the solution is often to start over and build from the ground up. EDRN, however, took diverse information and systems and connected them together using JPL's Object Oriented Data Technology (OODT) software. OODT won runner-up for NASA Software of the Year in 2003 by serving as a science data system framework allowing for systems to be quickly built and integrated using existing information as a way to improve analysis of distributed data. OODT is heavily used at NASA to build mission data systems, share data sets for planetary and earth science and to support advanced climate research.

The EDRN Knowledge System brings together the EDRN laboratories to pull information about biomarker, studies, specimens and results into an integrated enterprise for biomarker research. Figure 1 below shows the ERNE, the EDRN Resource Network Exchange (ERNE). ERNE was the first application built to support sharing of data for EDRN using OODT. Fifteen sites have been connected to share specimen information. In addition, JPL has worked across EDRN to support the capture, sharing, integration and analysis of EDRN data and to ensure that knowledge system provides the infrastructure needed to not only support sharing of data for single data objects such as specimens, but provide an entire knowledge system that semantically links all types of information together in order to transform EDRN into a true collaborative network with a deep knowledge-base of biomarker research. In establishing this infrastructure, the informatics team has adhered to a set of critical goals for ensuring the informatics platform scaled to meet the long-term objectives of EDRN and could continue to be expanded as a success story for informatics within the National Cancer Institute. These goals include (NOTE $4^{th}$ report here):

1. Defining an information model for describing the EDRN problem space;
2. enabling all components of the knowledge system to be distributed;
3. providing software interfaces for capture, discovery and access of data resources;
4. providing a secure transfer and distribution infrastructure to meet United States federal regulations for data sharing; and
5. providing an integrated portal environment across the distributed EDRN

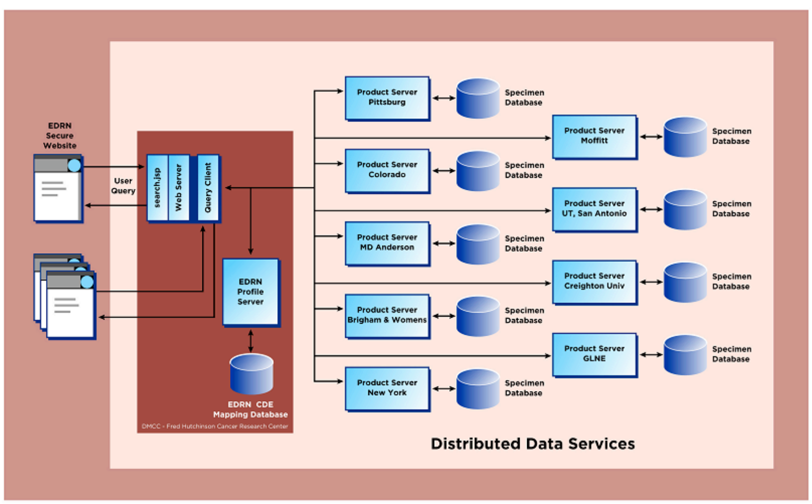


**Figure 1: EDRN Resource Network Exchange for Specimens**

While Figure 1 above shows the ERNE, our specimen sharing infrastructure, the EDRN has continued to expand the EDRN Knowledge System, as mentioned, to include eCAS, an infrastructure for warehousing biomarker data results, the biomarker database for capturing annotations, and VSIMS, a validation study infrastructure. All of this is tied together through a public portal that provides a semantic, "Google-like" search capability for finding and navigating results from EDRN biomarker studies.  Figure 2 shows the architecture model for the EDRN Knowledge Environment.

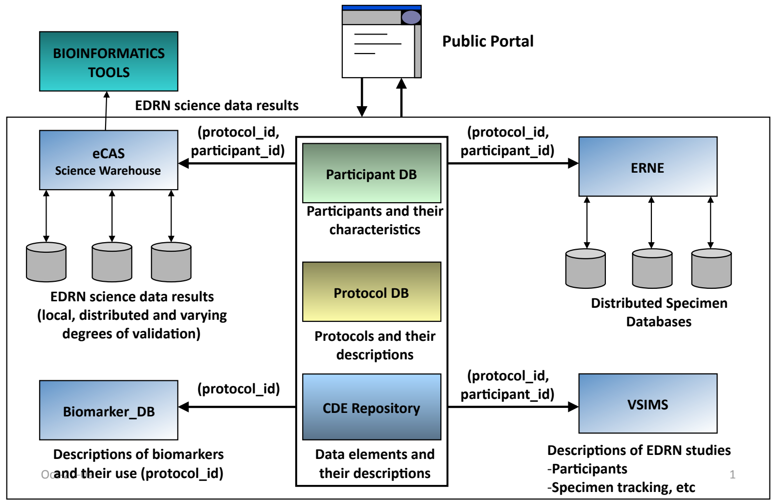


**Figure 2: EDRN Knowledge Environment Architecture**

## II. TECHNICAL PROGRESS

### 1. Developing the EDRN Informatics Architecture and Vision for Biomarker Research

Architectures, particularly in software research and development, are strategic and useful for providing a roadmap for the development and integration of a large-scale informatics system. JPL has played a key role in definition of the EDRN informatics architecture. For EDRN, the architecture describes the software components and services along with the associated biomarker data types that support the specific needs of biomarker research and specifically the EDRN researchers at the Biomarker Development Laboratories, Biomarker Reference Laboratories, and Clinical Epidemiology and Validation Centers, and the Data Management and Coordinating Center.

The EDRN informatics architecture is a highly distributed architecture that brings together a data and computational infrastructure of diverse groups and institutions to form the *EDRN Knowledge Environment* for early detection of cancer biomarkers.  EDRN, through the help of JPL, was one of the pioneers in defining and deploying this approach [10].  With the successful definition and implementation, it is now critical that EDRN begin to promote its architecture and ensure that it can be interoperable with other NCI and cancer research programs. For example, JPL on behalf of the EDRN, is working with the Canary Foundation to allow for data sharing with their informatics infrastructure.

### 2. Development of the core EDRN informatics infrastructure

During the 2001-2010 period, JPL was instrumental in helping to pioneer the use of distributed data system technologies to enable sharing of specimen, study and biomarker research results.

This infrastructure has helped to form what is known as the *EDRN Knowledge Environment.* A critical element of this was providing a software-computing infrastructure which allows for sharing of diverse databases across the EDRN. It is the objective of JPL in the next phase of EDRN to continue to ensure the underlying informatics infrastructure is in place, installed and connected at EDRN research laboratories, and is used to enable sharing of distributed information to allow researchers access to the wealth of biomarker data within the EDRN collaborative enterprise. This includes ensuring that best practices are used in applying modern information technologies and associated standards to allow for distributed access to the EDRN Knowledge Environment.

JPL has substantial background in developing such national infrastructures. First, JPL has developed a "data grid" middleware that enables sharing and computational services called the *Object Oriented Data Technology* (OODT) middleware [3]. This is a set of open source data grid components that allow for construction of distributed, data sharing systems around a set of distributed informatics services. OODT has been successfully integrated across the EDRN allowing for sharing of data. One of the most successful examples has been ERNE or the EDRN Resource Network Exchange. ERNE allows EDRN sites to share information about available specimens without changing their underling infrastructures. OODT, in this case, has been deployed to over fourteen EDRN sites. In addition to ERNE, OODT has served as the under-pinning informatics infrastructure for the EDRN Knowledge Environment allowing for sharing of scientific results and other information from biomarker validation studies.

Outside EDRN, OODT is highly leveraged by several projects at NASA where it serves as the infrastructure for planetary science data management and archiving for the Planetary Data System, the science data processing infrastructure for several earth science missions (Orbiting Carbon Observatory, NPP Sounder PEATE, SMAP, and QuikSCAT/SeaWinds), and the for data exchange to enable inter-comparison between satellite observations and climate forecast models for the climate research community. Figure 3 below shows the OODT middleware concept of sharing data and building virtual informatics systems.

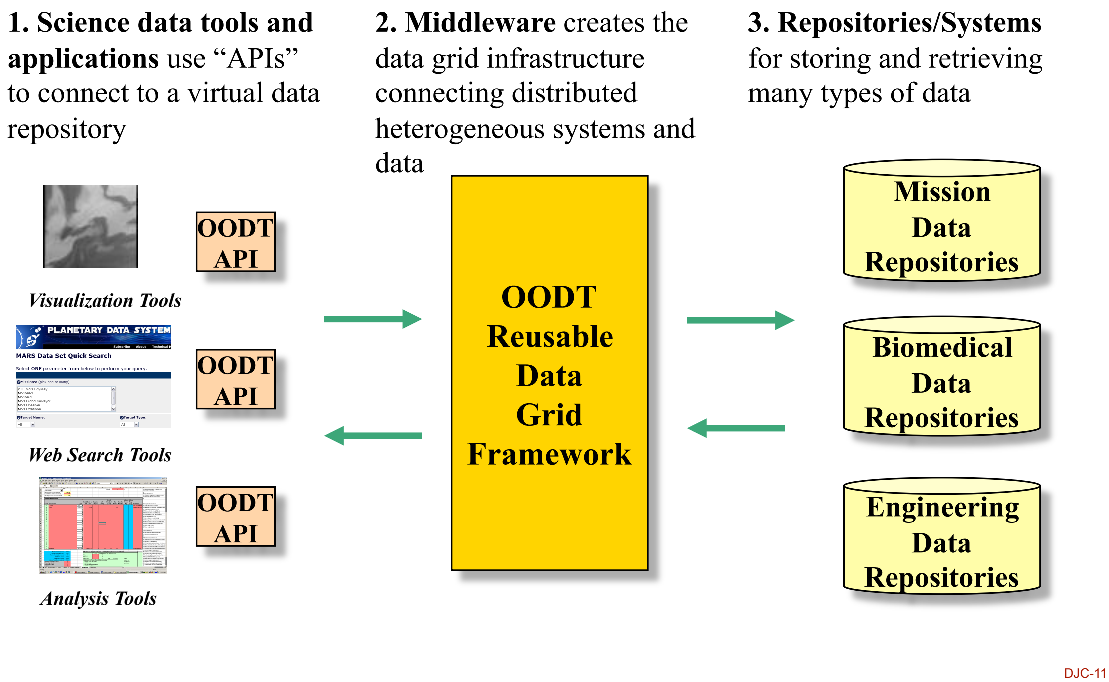


**Figure 3: OODT Middleware Concept**

## 3. Development and coordination of core biomarker databases

JPL has worked very closely with NCI, the DMCC and the investigators to develop a set of online databases to support biomarker research. These databases have been integrated into the EDRN Knowledge Environment and their access is served to the research community via the EDRN public portal. These databases include:

- EDRN Biomarker Database
- EDRN Catalog and Archive Service (eCAS)
- Validation Study Information Management System (VSIMS) and the EDRN Study Information System (eSIS)
- EDRN Resource Network Exchange (ERNE)

Figure 4 below shows these databases.

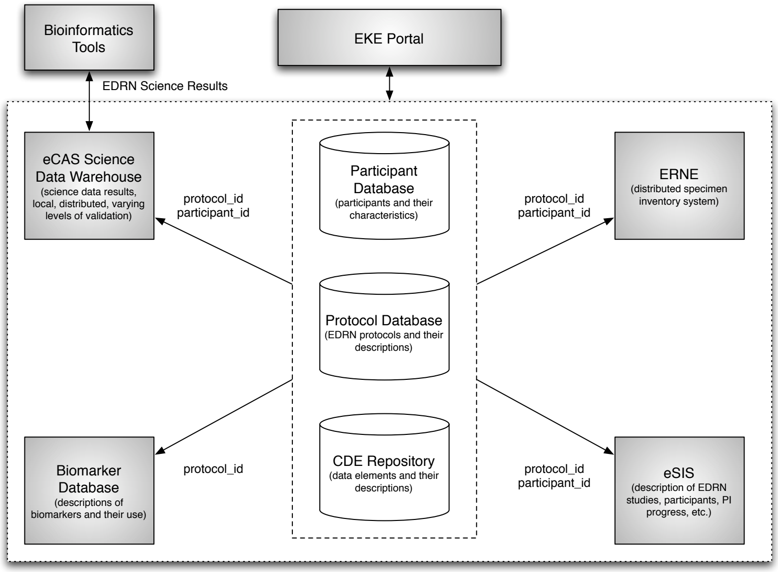


**Figure 4: EDRN Knowledge Environment with Key Databases**

The *Biomarker Database* is an EDRN resource that captures, integrates and annotates information regarding biomarkers. The primary focus has been the capture and annotation of biomarkers under study by the EDRN. The biomarker includes information about the biomarker itself, related resources, results from biomarker studies for specific organs, and related publications. While this information often exists in many different places, JPL has worked with the NCI, DMCC and other groups to "semantically" link this information into an integrated database. In addition, JPL has worked with the EDRN biomarker database curator at Dartmouth Medical School (K. Anton) to help capture this information. In support of these activities JPL has developed both a curation system as well as an extension of the EDRN public portal to allow users to search, view and access results from biomarker research within EDRN.

The *EDRN Catalog and Archive Service* provides an infrastructure for warehousing EDRN biomarker data results including both processed and unprocessed data sets. The infrastructure includes both the ability to curate the data by biocurators as well as access and view the data by the research community via the EDRN public portal. The eCAS implementation for EDRN directly leverages the EDRN ontology this allows for configuration of each data set and associated data product that is warehoused to be defined by a set of Common Data Elements (CDEs). It is critical to note that the definition for each data set can evolve as the science and technology evolve that captures the data. Each data set and associated data product are stored in eCAS along with their metadata descriptions. These are eventually exported to the public portal, as mentioned. In addition, JPL has been working closely with the Canary Foundation to allow for warehousing and sharing of data for collaborative projects with other groups. This allows users to bring up the EDRN portal, search for scientific data sets, and download those data sets regardless of whether their physical data is stored on EDRN hardware or within another data system environment. Like other components, eCAS includes the rich security model so that users are both authenticated and authorized within the system so they only see and access data for which they are granted permission.

The *Validation Study Information Management System* (VSIMS) and the *EDRN Study Information System* (eSIS) were developed by the EDRN DMCC to support validation studies information and coordination. They are critical to the overall EDRN Knowledge Environment and are therefore integrated with the rest of the system. Like other databases, the information is published to the EDRN portal for viewing. JPL and the DMCC have worked together to ensure that these databases are built from the EDRN Biomarker Ontology and that the information can be shared with the rest of the system. JPL deployed services at the DMCC to allow for sharing of this information with the portal as well as the Biomarker Database and eCAS since all of these need information about studies, sites, investigators and related publications.

One of EDRN's early informatics successes was deploying a national virtual specimen sharing system called ERNE, for *EDRN Resource Network Exchange* [14, 15]. ERNE enables applications to share biospecimen inventory annotations across EDRN. ERNE was successful because, while each system retained data in its native format, biospecimen objects that were shared were mapped and exchanged using the EDRN CDEs. Figure 5 below shows ERNE and the components. These components are distributed at several EDRN sites including University of Colorado, University of Pittsburgh, Dartmouth Medical Center, Creighton University, Duke University, University of California, San Diego, and more. The primary interface for ERNE is deployed at the Fred Hutchinson Cancer Center. Each site runs a set of OODT servers that enable it to share data.

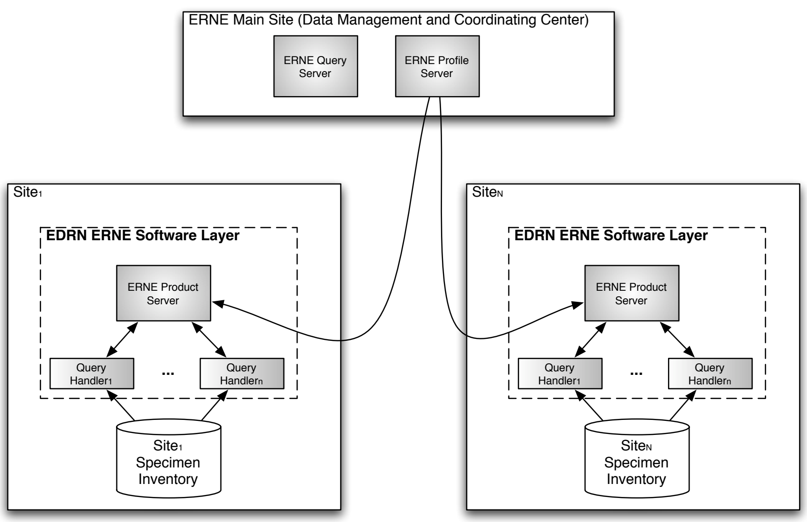


**Figure 5: ERNE System**

## 4. Development and coordination of an EDRN-wide Portal

The center-piece of the EDRN Knowledge Environment is the *EDRN Knowledge System Portal*. The portal serves dual purposes including release of public information as well as a portal to serve multiple stakeholders including the EDRN research community, the NCI staff, and the public. In addition to news, it contains significant scientific content that is all semantically linked providing a state-of-the-art interface for users to access and share results from EDRN research.

The public portal provides several features including

- News and information about the EDRN program
- Biomarker data and annotations
- EDRN site and investigator information
- EDRN study information
- Data sets from EDRN research
- EDRN Publications
- Google-like search to find information and results

Each of the EDRN applications described thus far, eCAS, ERNE, eSIS and the Biomarker Database, all publish content to the portal using the Resource Description Framework (RDF), a semantic web standard. RDF describes content (called "subjects") by first identifying them with URIs. For EDRN, the URIs to eCAS data are the URLs to the eCAS installation plus the path to the data provided by the eCAS server. For ERNE, the URIs are URLs to individual specimen records. For the Biomarker Database, URIs are URLs to single biomarkers tracked in the database. Figure 6 shows this concept.

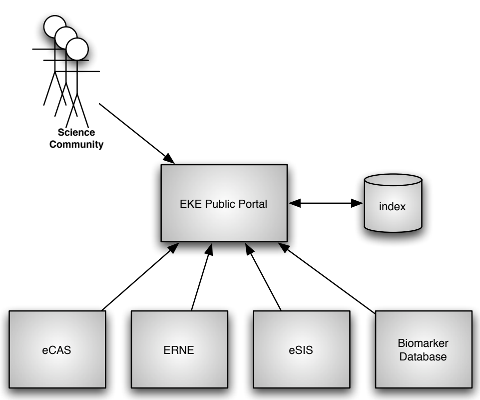


**Figure 6: Publishing Data to EDRN Portal**

JPL has built the portal using an open source technology known as Plone. Plone provides a content management infrastructure which allows content managers the ability to update and publish news and other information without making changes to the website. JPL has been one of the leaders in the use of the open source technology itself, presenting at several conferences on its unique use to support scientific data dissemination. JPL works with the DMCC and NCI to help them with the content management functions and to ensure that content for the portal is kept up to date.

In addition to news content, the latest feature of the portal is full integration of the biomarker data and science results. JPL has worked to integrate the content as mentioned above and to provide a rich set of capabilities for navigating and viewing that content. The portal provides the ability to search and find information using an advanced, semantic search with a free text search engine much like Google ™. The search results are organized into *facets* which allow users to navigate categories of information such as studies, biomarkers, science data, etc. Figure 7 below shows this view.

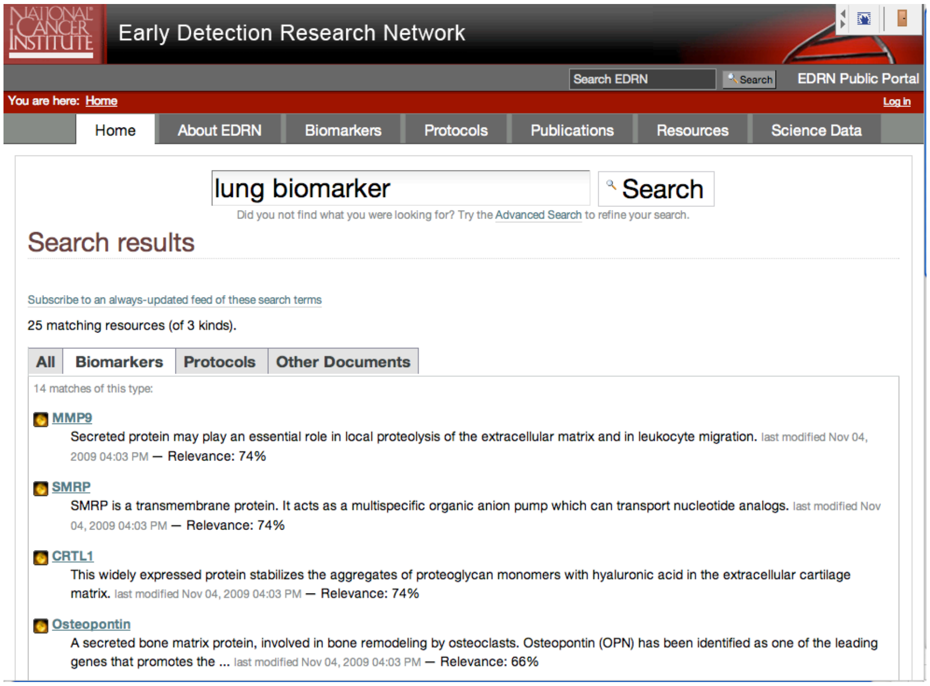


**Figure 7: EDRN Portal Biomarker Data Search with Facets**

JPL has worked with the National Cancer Institute to support hosting and sharing of the services and information. This includes working with Terpsys, a contractor for the Division of Cancer Prevention, to ensure that EDRN web services deployed at NCI under the cancer.gov domain, can integrate with other informatics services across EDRN.

## 5. Development of Biomarker Data Model and Informatics Standards

One of the major deliverables during the past period was development of an *EDRN Biomarker Ontology* that serves as a blueprint to describe the data, and it relationship, produced during biomarker research. An ontology is a modern mechanism for capturing and defining conceptual relationships and has been found to be highly effective in describing information produced for specific scientific domains. The EDRN ontology includes information about biomarkers, technologies, studies, results, publications, sites, and other information that forms a "model" for biomarker research. From a software system architecture perspective, the JPL approach is to put the intelligence in the ontology, rather than in software. This is because science is always evolving as new discoveries and approaches need to be supported by data analysis infrastructures. Therefore, the software architecture approach is to be ontology or model-driven where the data definitions are described in the model so that the underlying software support can dynamically evolve with the model to support new types of data, instruments, etc. It is our plan to make the EDRN ontology publically available so that it evolves as a standard for constructing informatics systems in early detection of cancer biomarkers. Figure 8 below shows a diagram of the ontology concept in EDRN.

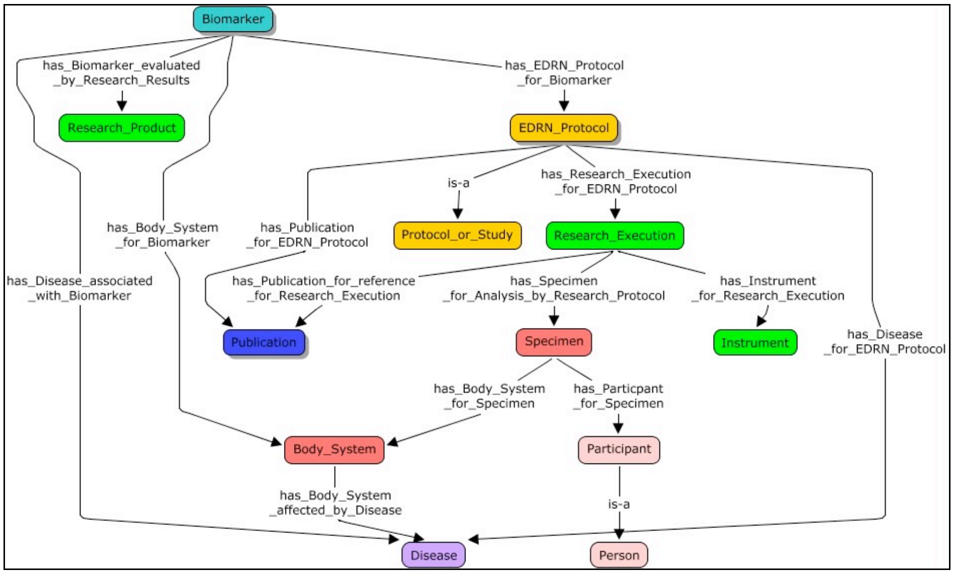

**Figure 8: EDRN Biomarker Ontology Conceptual Model**

In addition to the model, JPL has worked very closely with the DMCC and EDRN to define a structure for capturing and managing *EDRN Common Data Elements*. Common Data Elements (CDEs) are attributes defined in the model, but used specifically in the capture and management of information. For example, EDRN has defined a standard set of CDEs for capturing and managing information about biospecimens. These CDEs have been important for building an integrated system for sharing biospecimens. The CDEs are based on an international standard known as ISO/IEC 11179.

**6. Develop data sharing policies for biomarker research and collaboration across collaborative groups and among EDRN researchers**

Over the past few years, JPL has worked with the Data Sharing and Informatics Sub-committee within the EDRN and NCI to develop policies for data sharing. These policies include recommendations for the following:

- data release
- data sharing
- security
- standard processes for acquisition, storage, organization, retrieval and maintenance over time

The data release policies describe the period of time that a researcher has to release data for public use via eCAS. The data sharing policies describe the requirements and recommendations for sharing data within EDRN, which is compliant with the broader NIH policies. The security policy describes the multi-level security plan mentioned earlier and which groups have access to which type of biomarker data as biomarkers transition through phases of development. The processes mentioned above address mechanisms for managing the data throughout the lifecycle of a biomarker. Figure 9 below shows that lifecycle.

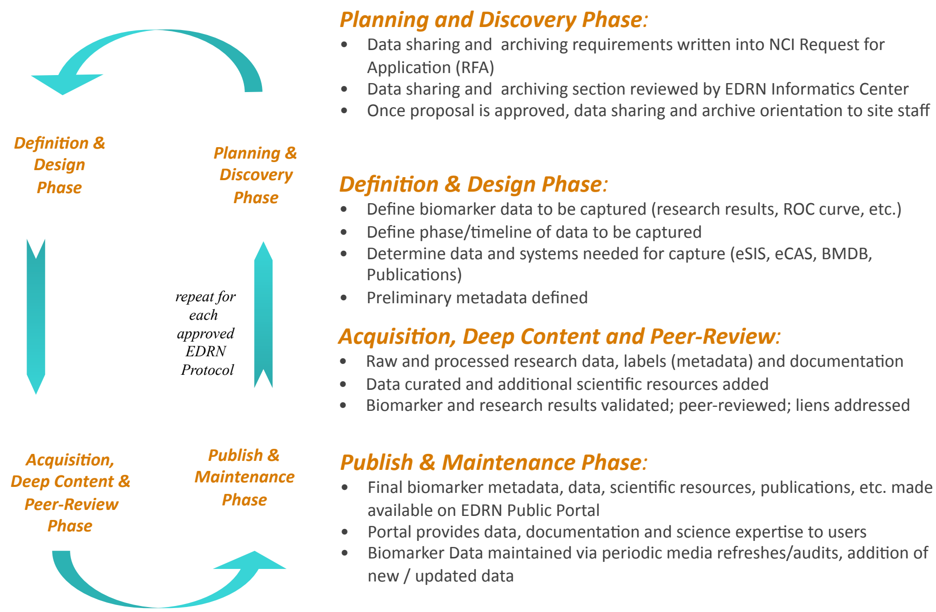


**Figure 9: Biomarker Data Lifecycle**

### 7: Provide curation of biomarker data results from EDRN studies

JPL and the Dartmouth Medical School worked very closely with NCI program directors and the DMCC to develop a curation process for the EDRN. This process includes developing a rich set of annotations of biomarkers and ensuring that the information is scientifically valid and useful. In addition, this was closely coordinated with tool and systems development to ensure that curation tools would enable the capture of information and that it could then be made available to EDRN research community and ultimately the public. This information includes biomarker studies and protocols, biomarker annotations, scientific data sets, publications and other related information.

## III. FUTURE PLANS

The EDRN informatics activities are leading-edge. They have helped the National Cancer Institute and biomedical research demonstrate the ability to support cross-institution collaboration through informatics. It has also shown the ability to improve data sharing and access by researchers both inside and outside the EDRN network. As such, the EDRN informatics infrastructure provides a robust foundation on which to continue to expand the program.

During the next period of performance, JPL plans to build on top of many of the capabilities including eCAS, ERNE, the biomarker database, and the public portal. These capabilities include improving the ability for researchers to directly process, capture and share their data both during the discovery and validation phases of a biomarker. In addition, JPL will work to capture the team project biomarkers and to develop a world-class database of biomarker information that will be useful to the community. Besides accessing biomarker information, the ERNE infrastructure will continue to be integrated into the EDRN public portal providing access both through the real-time data sharing infrastructure as well to annotated information about reference sets and other information.

One of the key capabilities that many researchers have accessed for is access to specific tools that support study management. JPL plans to actively work with the EDRN research community to adopt and develop, where needed, tools to support study management through a web-based infrastructure.

Finally, the infrastructure planned to be delivered through an open source model. This includes not only sharing the software, but actively publishing and distributing it through major open source infrastructures including the Apache Software Foundation and SourceForge.

## IV. EDRN-RELATED PUBLICATIONS

### Book Chapters


[1] D. Crichton, C. Mattmann, J. S. Hughes, S. Kelly, and A. Hart. A Multi-Disciplinary, Model-Driven, Distributed Science Data System Architecture. In Guide to e-Science: Next Generation Scientific Research and Discovery. X. Yang, L. L. Wang, W. Jie, eds. Spring Verlag, 2010.

[2] D. Crichton, C. Mattmann, M. Thornquist, J. S. Hughes, K. Anton. "Bioinformatics: Biomarkers of Early Detection." In *Translational Pathology of Early Cancer*. W. Grizzle, S. Srivastava, eds. IOS Press, 2010, To appear.

[3] D. Crichton, H. Kincaid, J.S. Hughes, S. Kelly, S. Srivastava, D. Johnsey. "Creating a National Virtual Knowledge Environment for Proteomics and Information Management". In *Informatics and Proteomics*. Marcel Dekker Publishers. December 2004.

[4] D. Crichton, J.S. Hughes and S. Kelly. "A Science Data System Architecture for Information Retrieval". In *Clustering and Information Retrieval.* Kluwer Academic Publishers. December 2003.


### Refereed Journals


[5] J. S. Hughes, D. Crichton and C. Mattmann. Ontology-Based Information Model Development for Science Information Reuse and Integration. International Transactions on Systems Science and Applications - Special Issue on Information Reuse in Databases and Data Mining, To appear, 2010.

[6] J. S. Hughes, D. Crichton, S. Kelly, C. Mattmann, T. Tran. "Intelligent Resource Discovery using Ontology-based Resource Profiles". *Data Science Journal*, Vol. 4, pp. 171-188, December 2005.

**Referred Conference Papers and Workshops**

[7] D. Crichton, C. Mattmann, E. Law, S. Hughes, S. Hardman, "Developing Software Product Lines for Science Data Systems", American Geophysical Union Meeting, San Francisco, December 2010.

[8] A. Hart, C. Mattmann, J. Tran, D. Crichton, H. Kincaid, J. S. Hughes, S. Kelly, K. Anton, D. Johnsey, C. Patriotis. "Enabling Effective Curation of Cancer Biomarker Research Data." To appear in Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS), Albuquerque, NM, August 3rd-4th, 2009.

[9] J. S. Hughes, D. Crichton and C. Mattmann. "Ontology-Based Information Model Development for Science Information Reuse and Integration." To appear in Proceedings of the 2009 IEEE International Conference on Information Reuse and Integration(IEEE IRI-09), Las Vegas, NV, August 10th-12th, 2009.

[10] W. Franklin, D. Crichton, M. Reid, C. Mattmann, A. Hart, D. Deng, P. Chesnut, B. Logue, J. Hayes, D. Stelling, M. Varella-Garcia, T. E. Kennedy, Y. E. Miller. "A Distributed Bronchial Mapping Software Tool for the Tracking of Cellular, Molecular and Imaging Results in the Central Airways." To appear in Proceedings of the 13th IASLC World Conference on Lung Cancer, San Francisco, CA, July 31st-August 4th, 2009.

[11] J. S. Hughes, D. Crichton, and C. Mattmann. "Scientific Digital Libraries, Interoperability, and Ontologies." In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009), pp. 399-400, Austin, TX, June 15-19, 2009.

[12] A. Hart, J. Tran, D. Crichton, K. Anton, H. Kincaid, S. Kelly, J.S. Hughes, C. Mattmann. "An Extensible Biomarker Curation Approach for Software Infrastructure for the Early Detection of Cancer." In *Proceedings of the IEEE International Conference on Health Informatics*. Porto, Portugal, January 14-17, 2009.

[13] Hart, D. Crichton, D. Johnsey, C. Mattmann, C. Patriotis, H. Kincaid, S. Srivastava, M. Thornquist. "A Web-based Data Management Infrastructure for Curation, Annotation and Dissemination of Biomarker Research results for the Early Detection of Cancer. " In Proceedings of the *5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.

[14] D. Crichton, M. Thornquist, S. Kelly, C. Mattmann, D. Johnsey, J. Dahlgren, D. Steling, G. Warnick, S. Reid, C. Edelstein, A. Hart, H. Kincaid. "A Distributed Informatics Knowledge Environment for Biomarker Research. " In Proceedings of the *5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.

[15] J. S. Hughes, D. Stelling, D.Crichton, C. Mattmann, G. Warnick, S. Reid. "An Information Model for Biomarker Research. " In Proceedings of the *5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.

[16] C. Mattmann, M. Khilkin, W. Rom, D. Crichton, S. Kelly, P. Rivera, J. Ko, B. Phalan, S. Sotero, E. Eylers. "A Reusable Web-based CAT (CT) scan data management system for temporally characterizing Solid Nodules and Ground Glass Opacities in Lung Cancer patients. " In Proceedings of the *5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.

[17] W. Franklin, D. Crichton, M. Reid, C. Mattmann, A.Hart, D. Deng, B. Logue, J. Hayes, D. Stelling. "A Distributed Biomarker Atlas for Lung Research aiding the Discovery and Early Detection of Cancer Biomarkers". In Proceedings of *the 5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.

[18] M. Khilkin, C. Mattmann, P. Rivera, J. Koh, B. Phalan, E. Eylers, S. Kelly, D. Crichton and W. N. Rom. "Integrating clinical, CT and PFT patient information in a database to determine a follow-up CT interval and the malignant potential of solid and ground glass pulmonary nodules." To appear in Proceedings of *American Thoracic Society (ATS),* Toronto, Ontario, Canada, May 16-21, 2008.

[19] M. Khilkin, C. Mattmann, P. Rivera, J. Koh, B. Phalan, E. Eylers, S. Kelly, D. Crichton and W. N. Rom. "An integrated clinical, CT, PFT database to better define an at risk population to screen for lung cancer." To appear in Proceedings of *American Thoracic Society (ATS),* Toronto, Ontario, Canada, May 16-21, 2008.

[20] C. Mattmann, V. Perrone, S. Kelly, D. Crichton, A. Finkelstein, and N. Medvidovic. A Reference Framework for Requirements and Architecture in Biomedical Grid Systems. In Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration(IEEE IRI-07), pp. 418-423, Las Vegas, NV, August 13-15, 2007.
[21]Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, and B. Bigbee. "A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer". Accepted for publication at the *2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam, the Netherlands, December 4th-6th, 2006.
[22] C. Mattmann, D. Crichton, N. Medvidovic and S. Hughes. "A Software Architecture-Based Framework for Highly Distributed and Data Intensive Scientific Applications". In Proceedings of the *28th International Conference on Software Engineering (ICSE06)*, pp. 721-730, Shanghai, China, May 20th-28th, 2006.
[23] C. Mattmann, D. Crichton, J. S. Hughes, S. Kelly and P. Ramirez. "Software Architecture for Large-scale, Distributed, Data-Intensive Systems". In Proceedings of the *4th Working IEEE/IFIP Conference on Software Architecture (WICSA-4)*. Oslo, Norway, June 12th-15th, 2004.
[24] D. Crichton, H. Kincaid, S. Kelly, S. Srivastava, D. Johnsey. "A National Data Grid Infrastructure for Sharing Biospecimens in Early Cancer Detection". *In Proceedings of the Digital Biology: the Emerging Paradigm*, Bethesda, MD. November 2003.
[25] D. Crichton, S. Srivastava and D. Johnsey. "A Virtual Bioinformatics Knowledge Environment for Early Cancer Detection". *7th World Multiconference on Systematics, Cybernetics and Informatics*. Orlando, Florida. July 2003.
[26] H. Kincaid, D. Crichton, M. Winget, S. Srivastava, D. Johnsey and M. Thornqusit. "A National Virtual Specimen Repository for Early Cancer Detection". *In Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems*. June 2003.
[27] D. Crichton, G. Downing, J.S. Hughes, H. Kincaid and S. Srivistava. "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network". *In Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems*. July 2001.

## V. REFERENCES

[1] D. Crichton, C. Mattmann, M. Thornquist, J. S. Hughes, K. Anton. "Bioinformatics: Biomarkers of Early Detection." In *Translational Pathology of Early Cancer*. W. Grizzle, S. Srivastava, eds. IOS Press, 2008, In Preparation.
[2] D. Crichton, H. Kincaid, J.S. Hughes, S. Kelly, S. Srivastava, D. Johnsey. "Creating a National Virtual Knowledge Environment for Proteomics and Information Management". In *Informatics and Proteomics*. Marcel Dekker Publishers. December 2004.
[3] D. Crichton, J.S. Hughes and S. Kelly. "A Science Data System Architecture for Information Retrieval". In *Clustering and Information Retrieval.* Kluwer Academic Publishers. December 2003.
[4] A. Hart, C. Mattmann, J. Tran, D. Crichton, H. Kincaid, J. S. Hughes, S. Kelly, K. Anton, D. Johnsey, C. Patriotis. "Enabling Effective Curation of Cancer Biomarker Research Data." To appear in Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS), Albuquerque, NM, August 3rd-4th, 2009.
[5] W. Franklin, D. Crichton, M. Reid, C. Mattmann, A. Hart, D. Deng, P. Chesnut, B. Logue, J. Hayes, D. Stelling, M. Varella-Garcia, T. E. Kennedy, Y. E. Miller. "A Distributed Bronchial Mapping Software Tool

[6] for the Tracking of Cellular, Molecular and Imaging Results in the Central Airways." To appear in Proceedings of the 13th IASLC World Conference on Lung Cancer, San Francisco, CA, July 31st-August 4th, 2009.
[7] A. Hart, J. Tran, D. Crichton, K. Anton, H. Kincaid, S. Kelly, J.S. Hughes, C. Mattmann. "An Extensible Biomarker Curation Approach for Software Infrastructure for the Early Detection of Cancer." To appear in *Proceedings of the IEEE International Conference on Health Informatics*. Porto, Portugal, January 14-17, 2009.
[8] D. Crichton, M. Thornquist, S. Kelly, C. Mattmann, D. Johnsey, J. Dahlgren, D. Steling, G. Warnick, S. Reid, C. Edelstein, A. Hart, H. Kincaid. "A Distributed Informatics Knowledge Environment for Biomarker Research. " In Proceedings of the *5th EDRN Scientific Workshop*, Bethesda, MD, March 17-19, 2008.
[9] M. Khilkin, C. Mattmann, P. Rivera, J. Koh, B. Phalan, E. Eylers, S. Kelly, D. Crichton and W. N. Rom. "Integrating clinical, CT and PFT patient information in a database to determine a follow-up CT interval and the malignant potential of solid and ground glass pulmonary nodules." To appear in Proceedings of *American Thoracic Society (ATS),* Toronto, Ontario, Canada, May 16-21, 2008.
[10] C. Mattmann, V. Perrone, S. Kelly, D. Crichton, A. Finkelstein, and N. Medvidovic. A Reference Framework for Requirements and Architecture in Biomedical Grid Systems. In Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration(IEEE IRI-07), pp. 418-423, Las Vegas, NV, August 13-15, 2007.
[11] D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, and B. Bigbee. "A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer". Accepted for publication at the *2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam, the Netherlands, December 4th-6th, 2006.
[12] D. Crichton, H. Kincaid, S. Kelly, S. Srivastava, D. Johnsey. "A National Data Grid Infrastructure for Sharing Biospecimens in Early Cancer Detection". *In Proceedings of the Digital Biology: the Emerging Paradigm*, Bethesda, MD. November 2003.
[13] R. Wetzel, D. Crichton, P. Ramirez, R. Kaptan, J. Fackler. "A National Informatics Infrastructure for Pediatric Intensive Care". *In Proceedings of the Digital Biology: the Emerging Paradigm*, Bethesda, MD. November 2003.
[14] D. Crichton, S. Srivastava and D. Johnsey. "A Virtual Bioinformatics Knowledge Environment for Early Cancer Detection". *7th World Multiconference on Systematics, Cybernetics and Informatics*. Orlando, Florida. July 2003.
[15] H. Kincaid, D. Crichton, M. Winget, S. Srivastava, D. Johnsey and M. Thornqusit. "A National Virtual Specimen Repository for Early Cancer Detection". *In Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems*. June 2003.
[16] D. Crichton, G. Downing, J.S. Hughes, H. Kincaid and S. Srivistava. "An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network". *In Proceedings of the 14th IEEE Symposium on Computer-Based Medical Systems*. July 2001.

## IV. ACKNOWLEDGEMENTS

The JPL Informatics Team includes Chris Mattmann, Heather Kincaid, Sean Kelly, Andrew Hart, Steve Hughes, John Tran, Rishi Verma and Dan Crichton. In addition, Kristen Anton from Dartmouth Medical School serves as the biocurator.

This work was conducted at the Jet Propulsion Laboratory, California Institute of Technology which is under contract to the National Aeronautics and Space Administration.